THE UNIVERSITY OF CHICAGO

# Stat 310 – Mihai Anitescu

## Lecture 8

# 8.4.1 OPTIMALITY CONDITIONS FOR EQUALITY CONSTRAINTS

# IFT for optimality conditions in the equality-only case

- Problem: $\boxed{(NLP)\ \min f(x) \text{ subject to } \mathrm{c}(x)=0;\ c:\mathbb{R}^n \to \mathbb{R}^m}$
- Assumptions:
  1. $x^*$ is a solution
  2. LICQ: $\nabla c(x)$ has full row rank.
- From LICQ: $\exists\, x^* = \left(\overbrace{x_{\mathcal{D}}^*}^{n-m}, \overbrace{x_{\mathcal{H}}^*}^{m}\right); \nabla c_{\mathcal{H}}(x^*) \in \mathbb{R}^{m\times m}; \nabla c_{\mathcal{H}}(x^*)$ invertible.
- From IFT:

$$\exists \mathcal{N}(x^*), \Psi(x_{\mathcal{D}}), \mathcal{N}(x_{\mathcal{D}}^*) \text{ such that } x \in \mathcal{N}(x^*) \cap \Omega \Leftrightarrow x_{\mathcal{H}} = \Psi(x_{\mathcal{D}})$$

- As a result $x^*$ is a solution of NLP iff $x_{\mathcal{D}}^*$ solves unconstrained problem: $\boxed{\min_{x_{\mathcal{D}}} f\left(x_{\mathcal{D}}, \Psi(x_{\mathcal{D}})\right)}$

–

# Properties of Mapping

- From IFT:

$$c\left(x_{\mathcal{D}},\Psi\left(x_{\mathcal{D}}\right)\right)=0 \Rightarrow \nabla_{x_{\mathcal{D}}} c\left(x_{\mathcal{D}},\Psi\left(x_{\mathcal{D}}\right)\right)+\nabla_{x_{\mathcal{H}}} c\left(x_{\mathcal{D}},\Psi\left(x_{\mathcal{D}}\right)\right)\nabla_{x_{\mathcal{D}}}\Psi\left(x_{\mathcal{D}}\right)=0$$

- Two important consequences

$$(1)\ \nabla_{x_{\mathcal{D}}}\Psi\left(x_{\mathcal{D}}\right)=-\left[\nabla_{x_{\mathcal{H}}} c\left(x_{\mathcal{D}},\Psi\left(x_{\mathcal{D}}\right)\right)\right]^{-1}\nabla_{x_{\mathcal{D}}} c\left(x_{\mathcal{D}},\Psi\left(x_{\mathcal{D}}\right)\right)$$

$$(2) Z=\begin{bmatrix} I_{n-m} \\ \nabla_{x_{\mathcal{D}}}\Psi\left(x_{\mathcal{D}}\right) \end{bmatrix} \Rightarrow \nabla c(x) Z=0 \Rightarrow \operatorname{Im}[Z]=\ker\left[\nabla c(x)\right]$$

# First-order optimality conditions

- Optimality of unconstrained optimization problem

$$\nabla_{x_{\mathcal{D}}} f\left(x_{\mathcal{D}}^*, \Psi\left(x_{\mathcal{D}}^*\right)\right) = 0 \Rightarrow \nabla_{x_{\mathcal{D}}} f\left(x_{\mathcal{D}}^*, \Psi\left(x_{\mathcal{D}}^*\right)\right) + \nabla_{x_{\mathcal{H}}} f\left(x_{\mathcal{D}}^*, \Psi\left(x_{\mathcal{D}}^*\right)\right) \nabla_{x_{\mathcal{D}}} \Psi\left(x_{\mathcal{D}}^*\right) = 0 \Rightarrow$$

$$\nabla_{x_{\mathcal{D}}} f\left(x_{\mathcal{D}}^*, \Psi\left(x_{\mathcal{D}}^*\right)\right) - \underbrace{\nabla_{x_{\mathcal{H}}} f\left(x_{\mathcal{D}}^*, \Psi\left(x_{\mathcal{D}}^*\right)\right)\left[\nabla_{x_{\mathcal{H}}} c\left(x_{\mathcal{D}}, \Psi\left(x_{\mathcal{D}}\right)\right)\right]^{-1}}_{\lambda^T} \nabla_{x_{\mathcal{D}}} c\left(x_{\mathcal{D}}, \Psi\left(x_{\mathcal{D}}\right)\right) = 0$$

- The definition of the Lagrange Multiplier Result in the first-order (Lagrange, KKT) conditions:

$$\left[\begin{array}{cc} \nabla_{x_{\mathcal{D}}} f\left(x_{\mathcal{D}}^*, \Psi\left(x_{\mathcal{D}}^*\right)\right) & \nabla_{x_{\mathcal{H}}} f\left(x_{\mathcal{D}}^*, \Psi\left(x_{\mathcal{D}}^*\right)\right) \end{array}\right] - \lambda^T \left[\begin{array}{cc} \nabla_{x_{\mathcal{D}}} c\left(x_{\mathcal{D}}, \Psi\left(x_{\mathcal{D}}\right)\right) & \nabla_{x_{\mathcal{H}}} c\left(x_{\mathcal{D}}^*, \Psi\left(x_{\mathcal{D}}^*\right)\right) \end{array}\right] = 0$$

$$\nabla f\left(x^*\right) - \lambda^T \nabla c\left(x^*\right) = 0$$

# A more abstract and general proof

- Optimality of unconstrained optimization problem

$$D_{x_{\mathcal{D}}} f\left(x_{\mathcal{D}}^*, \Psi\left(x_{\mathcal{D}}^*\right)\right)=0 \Rightarrow \nabla_{x_{\mathcal{D}}} f\left(x_{\mathcal{D}}^*, \Psi\left(x_{\mathcal{D}}^*\right)\right)+\nabla_{x_{\mathcal{H}}} f\left(x_{\mathcal{D}}^*, \Psi\left(x_{\mathcal{D}}^*\right)\right) \nabla_{x_{\mathcal{D}}} \Psi\left(x_{\mathcal{D}}^*\right)=0 \Rightarrow \nabla_x f\left(x^*\right) Z=0$$

- Using $\ker M \perp \operatorname{Im} M^T ; \dim(\ker M)+\dim\left(\operatorname{Im} M^T\right)=\text{nr cols M}$
- We obtain: $\nabla_x f\left(x^*\right) Z=0 \Rightarrow \nabla_x f\left(x^*\right)^T \in \ker\left(Z^T\right)=\operatorname{Im}\left[\nabla c\left(x^*\right)^T\right]$
- We thus obtain the optimality conditions:

$$\exists \lambda \in \mathbb{R}^m \text{ s.t. } \nabla_x f\left(x^*\right)^T=\nabla_x c\left(x^*\right)^T \lambda \Rightarrow \nabla_x f\left(x^*\right)-\lambda^T \nabla_x c\left(x^*\right)=0$$

# The Lagrangian

- Definition $\mathcal{L}(x,\lambda) = f(x) - \lambda^T c(x)$
- Its gradient $\nabla \mathcal{L}(x,\lambda) = \left[ \nabla f(x) - \lambda^T \nabla c(x), c(x)^T \right]$
- Its Hessian $\nabla^2 \mathcal{L}(x,\lambda) = \begin{bmatrix} \nabla^2_{xx} \mathcal{L}(x,\lambda) & \nabla c(x)^T \\ \nabla c(x) & 0 \end{bmatrix}$
- Where $\nabla^2_{xx} \mathcal{L}(x,\lambda) = \nabla^2_{xx} f(x,\lambda) - \sum_{i=1}^{m} \lambda_i \nabla^2_{xx} c_i(x,\lambda)$
- Optimality conditions: $\boxed{\nabla \mathcal{L}(x,\lambda) = 0}$

# Second-order conditions

- First, note that: $Z^T \nabla^2_{xx} L\left(x_{\mathcal{D}}, \Psi\left(x_{\mathcal{D}}\right)\right) Z = D^2_{x_{\mathcal{D}} x_{\mathcal{D}}} f\left(x_{\mathcal{D}}, \Psi\left(x_{\mathcal{D}}\right)\right) \succ = 0$
- Sketch of proof: total derivatives in $x_{\mathcal{D}}$ :

$$D_{x_{\mathcal{D}}} f\left(x_{\mathcal{D}}, \Psi\left(x_{\mathcal{D}}\right)\right) = \nabla_{x_{\mathcal{D}}} f\left(x_{\mathcal{D}}, \Psi\left(x_{\mathcal{D}}\right)\right) - \lambda\left(x_{\mathcal{D}}, \Psi\left(x_{\mathcal{D}}\right)\right)^T \nabla_{x_{\mathcal{D}}} c\left(x^*_{\mathcal{D}}, \Psi\left(x^*_{\mathcal{D}}\right)\right) =$$

$$\nabla_{x_{\mathcal{D}}} \mathcal{L}\left(\left(x_{\mathcal{D}}, \Psi\left(x_{\mathcal{D}}\right)\right), \lambda\left(x_{\mathcal{D}}, \Psi\left(x_{\mathcal{D}}\right)\right)\right);$$

$$\nabla_{x_{\mathcal{H}}} f\left(x^*_{\mathcal{D}}, \Psi\left(x^*_{\mathcal{D}}\right)\right) = \lambda\left(x_{\mathcal{D}}, \Psi\left(x_{\mathcal{D}}\right)\right)^T \nabla_{x_{\mathcal{H}}} c\left(x^*_{\mathcal{D}}, \Psi\left(x^*_{\mathcal{D}}\right)\right)$$

- Second derivatives:

$$D_{x_{\mathcal{D}} x_{\mathcal{D}}} f\left(x_{\mathcal{D}}, \Psi\left(x_{\mathcal{D}}\right)\right) = \nabla_{x_{\mathcal{D}}} f\left(x_{\mathcal{D}}, \Psi\left(x_{\mathcal{D}}\right)\right) - \lambda\left(x_{\mathcal{D}}, \Psi\left(x_{\mathcal{D}}\right)\right)^T \nabla_{x_{\mathcal{D}}} c\left(x_{\mathcal{D}}, \Psi\left(x_{\mathcal{D}}\right)\right) =$$

$$\nabla_{x_{\mathcal{D}} x_{\mathcal{D}}} \mathcal{L}\left(\left(x_{\mathcal{D}}, \Psi\left(x_{\mathcal{D}}\right)\right), \lambda\left(x_{\mathcal{D}}, \Psi\left(x_{\mathcal{D}}\right)\right)\right) + \nabla_{x_{\mathcal{D}}} \Psi\left(x_{\mathcal{D}}\right)^T \nabla_{x_{\mathcal{H}} x_{\mathcal{D}}} \mathcal{L}\left(\left(x_{\mathcal{D}}, \Psi\left(x_{\mathcal{D}}\right)\right), \lambda\left(x_{\mathcal{D}}, \Psi\left(x_{\mathcal{D}}\right)\right)\right)$$

$$-D_{\mathcal{D}}\left(\lambda\left(x_{\mathcal{D}}, \Psi\left(x_{\mathcal{D}}\right)\right)^T\right) \nabla_{x_{\mathcal{D}}} c\left(x_{\mathcal{D}}, \Psi\left(x_{\mathcal{D}}\right)\right)$$

# Computing Second-Order Derivatives

- Expressing the second derivatives of Lagrangian

$$\nabla_{x_{\mathcal{H}}} f\left(x_{\mathcal{D}}^{*}, \Psi\left(x_{\mathcal{D}}^{*}\right)\right) = \lambda\left(x_{\mathcal{D}}, \Psi\left(x_{\mathcal{D}}\right)\right)^{T} \nabla_{x_{\mathcal{H}}} c\left(x_{\mathcal{D}}, \Psi\left(x_{\mathcal{D}}\right)\right) \Rightarrow$$

$$D_{x_{\mathcal{D}}}\left[\lambda\left(x_{\mathcal{D}}, \Psi\left(x_{\mathcal{D}}\right)\right)^{T}\right] \nabla_{x_{\mathcal{H}}} c\left(x_{\mathcal{D}}, \Psi\left(x_{\mathcal{D}}\right)\right) = D_{x_{\mathcal{D}}}\left[\nabla_{x_{\mathcal{H}}} f\left(x_{\mathcal{D}}, \Psi\left(x_{\mathcal{D}}\right)\right) - \underbrace{\lambda\left(x_{\mathcal{D}}, \Psi\left(x_{\mathcal{D}}\right)\right)^{T}}_{inactive} \nabla_{x_{\mathcal{H}}} c\left(x_{\mathcal{D}}, \Psi\left(x_{\mathcal{D}}\right)\right)\right] =$$

$$D_{x_{\mathcal{D}}} \nabla_{x_{\mathcal{H}}} \mathcal{L}\left(\left(x_{\mathcal{D}}, \Psi\left(x_{\mathcal{D}}\right)\right), \underbrace{\lambda\left(x_{\mathcal{D}}, \Psi\left(x_{\mathcal{D}}\right)\right)^{T}}_{inactive}\right) = \nabla_{x_{\mathcal{D}}} \nabla_{x_{\mathcal{H}}} \mathcal{L}\left(\left(x_{\mathcal{D}}, \Psi\left(x_{\mathcal{D}}\right)\right), \lambda\left(x_{\mathcal{D}}, \Psi\left(x_{\mathcal{D}}\right)\right)^{T}\right) +$$

$$\nabla_{x_{\mathcal{D}}} \Psi\left(x_{\mathcal{D}}\right)^{T} \nabla_{x_{\mathcal{H}}} \nabla_{x_{\mathcal{H}}} \mathcal{L}\left(\left(x_{\mathcal{D}}, \Psi\left(x_{\mathcal{D}}\right)\right), \lambda\left(x_{\mathcal{D}}, \Psi\left(x_{\mathcal{D}}\right)\right)^{T}\right)$$

- Solve for total derivative of multiplier and replace conclusion follows.

# Summary: Necessary Optimality Conditions

- Summary:

$$\boxed{\nabla \mathcal{L}\left(x^*,\lambda^*\right)=0;\ Z^T \nabla_{xx}^2 L\left(x_{\mathcal{D}}^*,\Psi\left(x_{\mathcal{D}}^*\right)\right) Z \succ= 0}$$

- Rephrase first order:

$$\nabla_x \mathcal{L}\left(x^*,\lambda^*\right)=0$$

- Rephrase second order necessary conditions.

$$\nabla_x c\left(x^*\right) w = 0 \Rightarrow w^T \nabla_{xx}^2 \mathcal{L}\left(x^*,\lambda^*\right) w \geq 0$$

# Sufficient Optimality Conditions

- The point is a local minimum if LICQ and the following holds:

$$(1)\nabla_x \mathcal{L}\left(x^*,\lambda^*\right)=0;\ (2)\nabla_x c\left(x^*\right)w=0 \Rightarrow \exists\sigma>0\ w^T\nabla_{xx}^2\mathcal{L}\left(x^*,\lambda^*\right)w \geq \sigma\|w\|^2$$

- Proof: By IFT, there is a change of variables such that

$$u\in\mathcal{N}(0)\subset\mathbb{R}^{n-n_c}u \leftrightarrow x(u);\ \tilde{x}\in\mathcal{N}\left(x^*\right), c(\tilde{x})=0 \Leftrightarrow \exists\tilde{u}\in\mathcal{N}(0);\ \tilde{x}=x(\tilde{u})$$

$$\nabla_x c\left(x^*\right)\nabla_u x(\tilde{u})\Big|_{\tilde{u}=0}=0;\quad Z=\nabla_u x(\tilde{u})$$

- The original problem can be phrased as

$$\min_u f\left(x(u)\right)$$

# Sufficient Optimality Conditions

- We can now piggy back on theory of unconstrained optimization, noting that.

$$\nabla_u f\left(x(u)\right)\Big|_{u=0} = \nabla_x \mathcal{L}\left(x^*,\lambda^*\right) = 0;$$

$$\nabla^2_{uu} f\left(x(u)\right)\Big|_{u=0} = Z^T \nabla^2_{xx} \mathcal{L}\left(x^*,\lambda^*\right) Z \succ 0;\, Z = \nabla_u x(u)$$

- Then from theory of unconstrained optimization

we have a local isolated minimum at 0 and thus the original problem at $x^*$. (following the local isomorphism above)

# Another Essential Consequence

- If LICQ+ second-order conditions hold at the solution $x^*$ , then the following matrix must be nonsingular (EXPAND).

$$\begin{bmatrix} \nabla^2_{xx}\mathcal{L}\left(x^*,\lambda^*\right) & \nabla_x c\left(x^*\right) \\ \nabla^T_x c\left(x^*\right) & 0 \end{bmatrix}$$

- The system of nonlinear equations has an invertible Jacobian,

$$\begin{bmatrix} \nabla_x\mathcal{L}\left(x^*,\lambda^*\right) \\ c\left(x^*\right) \end{bmatrix} = 0$$

# 8.4.2 FIRST-ORDER OPTIMALITY CONDITIONS FOR MIXED EQ AND INEQ CONSTRAINTS

# The Lagrangian

- Even in the general case, it has the same expression

$$\mathcal{L}(x) = f(x) - \sum_{i \in \mathcal{E} \cup \mathcal{A}} \lambda_i c_i(x)$$

# First-Order Optimality Condition Theorem

*Suppose that $x^*$ is a local solution of (12.1), that the functions $f$ and $c_i$ in (12.1) are continuously differentiable, and that the LICQ holds at $x^*$. Then there is a Lagrange multiplier vector $\lambda^*$, with components $\lambda_i^*$, $i \in \mathcal{E} \cup \mathcal{I}$, such that the following conditions are satisfied at $(x^*, \lambda^*)$*

$$\nabla_x \mathcal{L}(x^*, \lambda^*) = 0, \tag{12.34a}$$

$$c_i(x^*) = 0, \quad \text{for all } i \in \mathcal{E}, \tag{12.34b}$$

$$c_i(x^*) \geq 0, \quad \text{for all } i \in \mathcal{I}, \tag{12.34c}$$

$$\lambda_i^* \geq 0, \quad \text{for all } i \in \mathcal{I}, \tag{12.34d}$$

$$\lambda_i^* c_i(x^*) = 0, \quad \text{for all } i \in \mathcal{E} \cup \mathcal{I}. \tag{12.34e}$$

Equivalent Form:

$$\nabla f\left(x^*\right) - \lambda^T_{\mathcal{A}\left(x^*\right)} \nabla c_{\mathcal{A}\left(x^*\right)}\left(x^*\right) = 0 \Rightarrow \text{Multipliers are unique !!}$$

# Sketch of the Proof

- If $x^*$ is a solution of the original problem, it is also a solution of the problem.

$$\min f(x) \text{ subject to } c_{\mathcal{A}(x^*)}(x) = 0$$

- From the optimality conditions of the problem with equality constraints, we must have (since LICQ holds)

$$\exists \{\lambda_i\}_{i \in \mathcal{A}(x^*)} \quad \text{such that} \quad \nabla f(x^*) - \sum_{i \in \mathcal{A}(x^*)} \lambda_i \nabla c_i(x^*) = 0$$

- But I cannot yet tell by this argument $\lambda_i \geq 0$

# Sketch of the Proof: The sign of the multiplier

- Assume now one multiplier has the "wrong" sign. That is

$$j \in \mathcal{A}(x^*) \cap \mathcal{I}, \quad \lambda_j < 0$$

- Since LICQ holds, we can construct a feasible path that "takes off" from that constraint (inactive constraints do not matter locally)
- $$c_{\mathcal{A}(x^*)}(\tilde{x}(t)) = te_j \Rightarrow \tilde{x}(t) \in \Omega \quad \text{Define } b = \frac{d}{dt}\tilde{x}(t)_{t=0} \Rightarrow \nabla c_{\mathcal{A}(x)} b = e_j$$

$$\frac{d}{dt} f(\tilde{x}(t))_{t=0} = \nabla f(x^*)^T b = \lambda_{c_{\mathcal{A}(x)}}^T \nabla c_{\mathcal{A}(x)} b = \lambda_j < 0 \quad \Rightarrow$$

$$\exists t_1 > 0, \quad f(\tilde{x}(t_1)) < f(\tilde{x}(0)) = f(x^*), \quad \text{CONTRADICTION!!}$$

# Strict Complementarity

- It is a notion that makes the problem look "almost" like an equality.

**Definition 12.5** (Strict Complementarity).

*Given a local solution $x^*$ of (12.1) and a vector $\lambda^*$ satisfying (12.34), we say that the* strict complementarity condition *holds if* exactly one *of $\lambda_i^*$ and $c_i(x^*)$ is zero for each index $i \in \mathcal{I}$. In other words, we have that $\lambda_i^* > 0$ for each $i \in \mathcal{I} \cap \mathcal{A}(x^*)$.*

# 8.5 SECOND-ORDER CONDITIONS

# Critical Cone

- The subset of the tangent space, where the objective function does not vary to first-order.
- The book definition.

$$\mathcal{C}(x^*, \lambda^*) = \{w \in \mathcal{F}(x^*) \mid \nabla c_i(x^*)^T w = 0, \text{ all } i \in \mathcal{A}(x^*) \cap \mathcal{I} \text{ with } \lambda_i^* > 0\}.$$

- An even simpler equivalent definition.

$$\mathcal{C}\left(x^*, \lambda^*\right) = \left\{w \in T_\Omega\left(x^*\right) \middle| \nabla f\left(x^*\right)^T w = 0\right\}$$

# Rephrasing of the Critical Cone

- By investigating the definition

$$w \in \mathcal{C}\left(x^*,\lambda^*\right) \Leftrightarrow \begin{cases} \nabla c_i\left(x^*\right)^T w = 0 & i \in \mathcal{E} \\ \nabla c_i\left(x^*\right)^T w = 0 & i \in \mathcal{A}\left(x^*\right) \cap \mathcal{I} \quad \lambda_i^* > 0 \\ \nabla c_i\left(x^*\right)^T w \geq 0 & i \in \mathcal{A}\left(x^*\right) \cap \mathcal{I} \quad \lambda_i^* = 0 \end{cases}$$

- In the case where strict complementarity holds, the cones has a MUCH simplex expression.

$$w \in \mathcal{C}\left(x^*,\lambda^*\right) \Leftrightarrow \nabla c_i\left(x^*\right) w = 0 \ \forall\, i \in \mathcal{A}\left(x^*\right)$$

# Statement of the Second-Order Conditions

**Theorem 12.5** (Second-Order Necessary Conditions).

*Suppose that $x^*$ is a local solution of (12.1) and that the LICQ condition is satisfied. Let $\lambda^*$ be the Lagrange multiplier vector for which the KKT conditions (12.34) are satisfied. Then*

$$w^T \nabla^2_{xx} \mathcal{L}(x^*, \lambda^*) w \geq 0, \quad \text{for all } w \in \mathcal{C}(x^*, \lambda^*). \tag{12.57}$$

- How to prove this? In the case of Strict Complementarity the critical cone is the same as the problem constrained with equalities on active index.
- Result follows from equality-only case.

# Statement of second-order sufficient conditions

**Theorem 12.6** (Second-Order Sufficient Conditions).

*Suppose that for some feasible point $x^* \in \mathbb{R}^n$ there is a Lagrange multiplier vector $\lambda^*$ such that the KKT conditions (12.34) are satisfied. Suppose also that*

$$w^T \nabla^2_{xx} \mathcal{L}(x^*, \lambda^*) w > 0, \quad \text{for all } w \in \mathcal{C}(x^*, \lambda^*), w \neq 0. \tag{12.65}$$

*Then $x^*$ is a strict local solution for (12.1).*

- How do we prove this? In the case of strict complementarity again from reduction to the equality case.

$$x^* = \arg\min_x f(x) \text{ subject to } c_{\mathcal{A}}(x) = 0$$

# How to derive those conditions in the other case?

- Use the slacks to reduce the problem to one with equality constraints.

$$
\begin{array}{llll}
\min_{x\in\mathbb{R}^n, z\in\mathbb{R}^{n_I},} & f(x) & & \\
s.t. & c_E(x) & = 0 & \\
 & \left[c_I(x)\right]_j - z_j^2 & = 0 & j = 1,2,\ldots n_I
\end{array}
$$

- Then, apply the conditions for equality constraints.
- I will assign it as homework.

# Summary: Why should I care about Lagrange Multipliers?

- Because it makes the optimization problem in principle equivalent to a nonlinear equation.

$$\begin{bmatrix} \nabla_x \mathcal{L}\left(x^*,\lambda^*\right) \\ c_{\mathcal{A}}\left(x^*\right) \end{bmatrix} = 0; \ \det \begin{bmatrix} \nabla_{xx}^2 \mathcal{L}\left(x^*,\lambda^*\right) & \nabla_x c_{\mathcal{A}}\left(x^*\right) \\ \nabla_x^T c_{\mathcal{A}}\left(x^*\right) & 0 \end{bmatrix} \neq 0$$

- I can use concepts from nonlinear equations such as Newton's for the algorithmics.

THE UNIVERSITY OF CHICAGO

# Section 9

## Fundamentals of Algorithms for Constrained Optimization

# 9.1 TYPES OF CONSTRAINED OPTIMIZATION ALGORITHMS

# Types of Optimization Algorithms

- All of the algorithms solve iteratively a simpler problem.
  - Penalty and Augmented Lagrangian Methods.
  - Sequential Quadratic Programming.
  - Interior-point Methods.
- The approach follows the usual divide-and-conquer approach:
    - Constrained Optimization-
    - Unconstrained Optimization
    - Nonlinear Equations
    - Linear Equations

# Quadratic Programming Problems

- Algorithms for such problems are interested to explore because
  - 1. Their structure can be efficiently exploited.
  - 2. They form the basis for other algorithms, such as augmented Lagrangian and Sequential quadratic programming problems.

$$
\begin{aligned}
\min_{x} \quad & q(x) = \tfrac{1}{2}x^T G x + x^T c \\
\text{subject to} \quad & a_i^T x = b_i, \qquad i \in \mathcal{E}, \\
& a_i^T x \geq b_i, \qquad i \in \mathcal{I},
\end{aligned}
$$

# Penalty Methods

- Idea: Replace the constraints by a penalty term.
- Inexact penalties: parameter driven to infinity to recover solution. Example:

$$x^* = \arg\min f(x) \text{ subject to } c(x) = 0 \Leftrightarrow$$

$$x^\mu = \arg\min \boxed{f(x) + \frac{\mu}{2}\sum_{i\in\mathcal{E}} c_i^2(x)};\ x^* = \lim_{\mu\to\infty} x^\mu = x^*$$

Solve with unconstrained optimization

- Exact but nonsmooth penalty – the penalty parameter can stay finite.

$$x^* = \arg\min f(x) \text{ subject to } c(x) = 0 \Leftrightarrow x^* = \arg\min f(x) + \mu\sum_{i\in\mathcal{E}} |c_i(x)|;\ \mu \geq \mu_0$$

# Augmented Lagrangian Methods

- Mix the Lagrangian point of view with a penalty point of view.

$$x^* = \arg\min f(x) \text{ subject to } \mathrm{c}(x) = 0 \Leftrightarrow$$

$$x^{\mu,\lambda} = \arg\min f(x) - \sum_{i\in\mathcal{E}} \lambda_i c_i(x) + \frac{\mu}{2}\sum_{i\in\mathcal{E}} c_i^2(x) \Rightarrow$$

$$x^* = \lim_{\lambda\to\lambda^*} x^{\mu,\lambda} \text{ for some } \mu \geq \mu_0 > 0$$

# Sequential Quadratic Programming Algorithms

- Solve successively Quadratic Programs.

$$
\begin{aligned}
\min_p \quad & \frac{1}{2} p^T B_k p + \nabla f(x_k) \\
\text{subject to} \quad & \nabla c_i(x_k) d + c_i(x_k) = 0 \quad i \in \mathcal{E} \\
& \nabla c_i(x_k) d + c_i(x_k) \geq 0 \quad i \in \mathcal{I}
\end{aligned}
$$

- It is the analogous of Newton's method for the case of constraints if $B_k = \nabla^2_{xx} \mathcal{L}(x_k, \lambda_k)$
- But how do you solve the subproblem? It is possible with extensions of simplex which I do not cover.
- An option is BFGS which makes it convex.

# Interior Point Methods

- Reduce the inequality constraints with a barrier

$$\begin{array}{lll} \min_{x,s} & f(x) - \mu \sum_{i=1}^{m} \log s_i & \\ \text{subject to} & c_i(x) = 0 & i \in \mathcal{E} \\ & c_i(x) - s_i = 0 & i \in \mathcal{I} \end{array}$$

- An alternative, is use a penalty as well:

$$\min_x f(x) - \mu \sum_{i \in \mathcal{I}} \log s_i + \frac{1}{2\mu} \sum_{i \in \mathcal{I}} \left(c_i(x) - s\right)^2 + \frac{1}{2\mu} \sum_{i \in \mathcal{E}} \left(c_i(x)\right)^2$$

- And I can solve it as a sequence of unconstrained problems!

# 9.2 MERIT FUNCTIONS AND FILTERS

# Feasible algorithms

- If I can afford to maintain feasibility at all steps, then I just monitor decrease in objective function.
- I accept a point if I have enough descent.
- But this works only for very particular constraints, such as linear constraints or bound constraints (and we will use it).
- Algorithms that do that are called feasible algorithms.

# Infeasible algorithms

- But, sometimes it is VERY HARD to enforce feasibility at all steps (e.g. nonlinear equality constraints).
- And I need feasibility only in the limit; so there is benefit to allow algorithms to move on the outside of the feasible set.
- But then, how do I measure progress since I have two, apparently contradictory requirements:
  – Reduce infeasibility (e.g. $\sum_{i \in \mathcal{E}} |c_i(x)| + \sum_{i \in \mathcal{I}} \max\{-c_i(x), 0\}$ )
  – Reduce objective function.
  – It has a multiobjective optimization nature!

# 9.2.1 MERIT FUNCTIONS

# Merit function

- One idea also from multiobjective optimization: minimize a weighted combination of the 2 criteria.

$$\phi(x) = w_1 f(x) + w_2 \left[ \sum_{i\in\mathcal{E}} |c_i(x)| + \sum_{i\in\mathcal{I}} \max\{-c_i(x), 0\} \right]; \quad w_1, w_2 > 0$$

- But I can scale it so that the weight of the objective is 1.
- In that case, the weight of the infeasibility measure is called "penalty parameter".
- I can monitor progress by ensuring that $\phi(x)$ decreases, as in unconstrained optimization.

# Nonsmooth Penalty Merit Functions

$$\phi_1(x;\mu) = f(x) + \mu \sum_{i \in \mathcal{E}} |c_i(x)| + \mu \sum_{i \in \mathcal{I}} [c_i(x)]^-, \qquad [z]^- = \max\{0, -z\}.$$

Penalty parameter

- It is called the l1 merit function.
- Sometimes, they can be even EXACT.

**Definition 15.1** (Exact Merit Function).

*A merit function $\phi(x;\mu)$ is* exact *if there is a positive scalar $\mu^*$ such that for any $\mu > \mu^*$, any local solution of the nonlinear programming problem (15.1) is a local minimizer of $\phi(x;\mu)$.*

We show in Theorem 17.3 that, under certain assumptions, the $\ell_1$ merit function $\phi_1(x;\mu)$ is exact and that the threshold value $\mu^*$ is given by

$$\mu^* = \max\{|\lambda_i^*|,\ i \in \mathcal{E} \cup \mathcal{I}\},$$

# Smooth and Exact Penalty Functions

- Excellent convergence properties, but very expensive to compute.
- Fletcher's augmented Lagrangian:

$$\phi_F(x; \mu) = f(x) - \lambda(x)^T c(x) + \tfrac{1}{2}\mu \sum_{i \in \mathcal{E}} c_i(x)^2,$$

$$\lambda(x) = [A(x)A(x)^T]^{-1} A(x) \nabla f(x).$$

- It is both smooth and exact, but perhaps impractical due to the linear solve.

# Augmented Lagrangian

- Smooth, but inexact.

$$\phi(x) = f(x) - \sum_{i \in \mathcal{E}} \lambda_i c_i(x) + \frac{\mu}{2} \sum_{i \in \mathcal{E}} c_i^2(x) \Rightarrow$$

- An update of the Lagrange Multiplier is needed.
- We will not uses it, except with Augmented Lagrangian methods themselves.

# Line-search (Armijo) for Nonsmooth Merit Functions

$$\phi_1(x;\mu) = f(x) + \mu \sum_{i\in\mathcal{E}} |c_i(x)| + \mu \sum_{i\in\mathcal{I}} [c_i(x)]^-,$$

- How do we carry out the "progress search"?
- That is the line search or the sufficient reduction in trust region?
- In the unconstrained case, we had

$$f(x_k) - f(x_k + \beta^m d_k) \geq -\rho\beta^m \nabla f(x_k)^T d_k; \quad 0 < \beta < 1, 0 < \rho < 0.5$$

- But we cannot use this anymore, since the function is not differentiable.

# Directional Derivatives of Nonsmooth Merit Function

$$\phi_1(x;\mu) = f(x) + \mu \sum_{i\in\mathcal{E}} |c_i(x)| + \mu \sum_{i\in\mathcal{I}} [c_i(x)]^-,$$

- Nevertheless, the function has a directional derivative (follows from properties of max function). EXPAND

$$D(\phi(x,\mu);p) = \lim_{t\to 0, t>0} \frac{\phi(x+tp,\mu)-\phi(x,\mu)}{t}; \quad D(\max\{f_1,f_2\},p) = \max\{\nabla f_1 p, \nabla f_1 p\}$$

- Line Search: $\phi(x_k,\mu)-\phi(x_k+\beta^m p_k,\mu) \geq -\rho\beta^m D(\phi(x_k,\mu),p_k);$
- Trust Region

$$\phi(x_k,\mu)-\phi(x_k+\beta^m p_k,\mu) \geq -\eta_1(m(0)-m(p_k));$$
$$0<\eta_1<0.5$$

# And …. How do I choose the penalty parameter?

- VERY tricky issue, highly dependent on the penalty function used.
- For the l1 function, guideline is:

$$\mu^* = \max\{|\lambda_i^*|,\ i \in \mathcal{E} \cup \mathcal{I}\},$$

- But almost always adaptive. Criterion: If optimality gets ahead of feasibility, make penalty parameter more stringent.
- E.g l1 function: the max of current value of multipliers plus safety factor (EXPAND)

–

# 9.2.2 FILTER APPROACHES

# Principles of filters

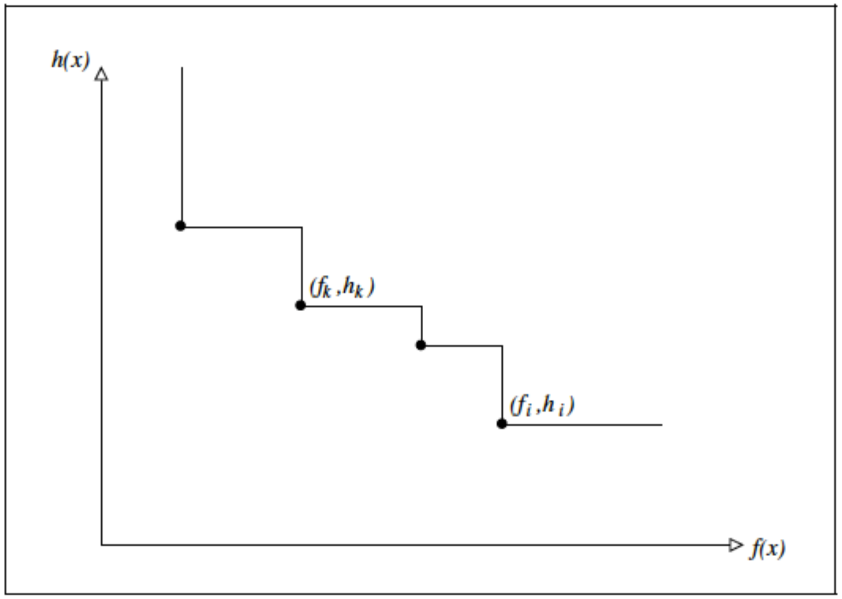
- Originates in the multiobjective optimization philosophy: objective and infeasibility

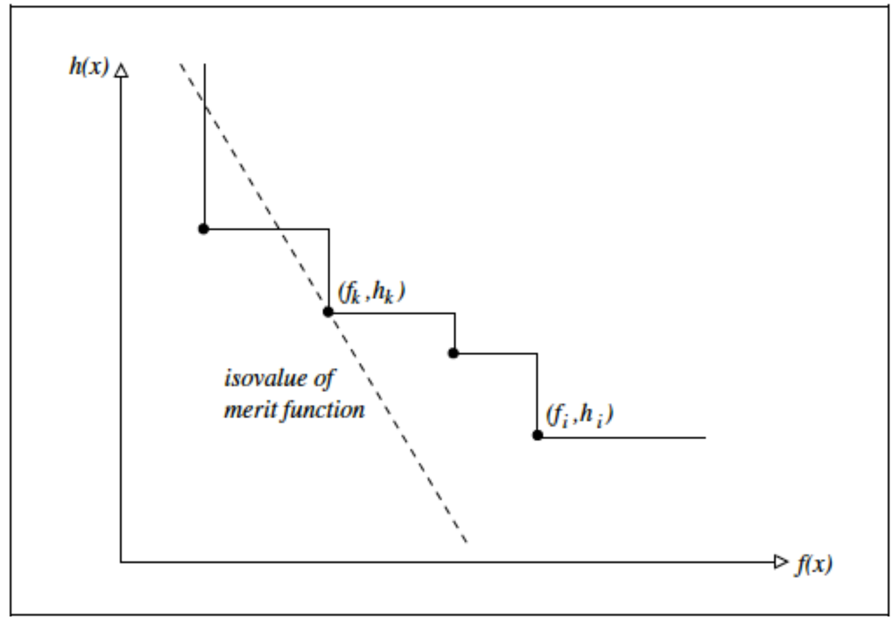
$$h(x) = \sum_{i \in \mathcal{E}} |c_i(x)| + \sum_{i \in \mathcal{I}} [c_i(x)]^-,$$

- The problem becomes:

$$\min_x f(x) \quad \text{and} \quad \min_x h(x).$$

# The Filter approach




**Definition 15.2.**

(a) *A pair* $(f_k, h_k)$ *is said to dominate another pair* $(f_l, h_l)$ *if both* $f_k \le f_l$ *and* $h_k \le h_l$.

(b) *A* filter *is a list of pairs* $(f_l, h_l)$ *such that no pair dominates any other.*

(c) *An iterate* $x_k$ *is said to be acceptable to the filter if* $(f_k, h_k)$ *is not dominated by any pair in the filter.*

# Some Refinements

- Like in the line search approach, I cannot accept EVERY decrease since I may never converge.
- Modification:

A trial iterate $x^+$ is acceptable to the filter if, for all pairs $(f_j, h_j)$ in the filter, we have that

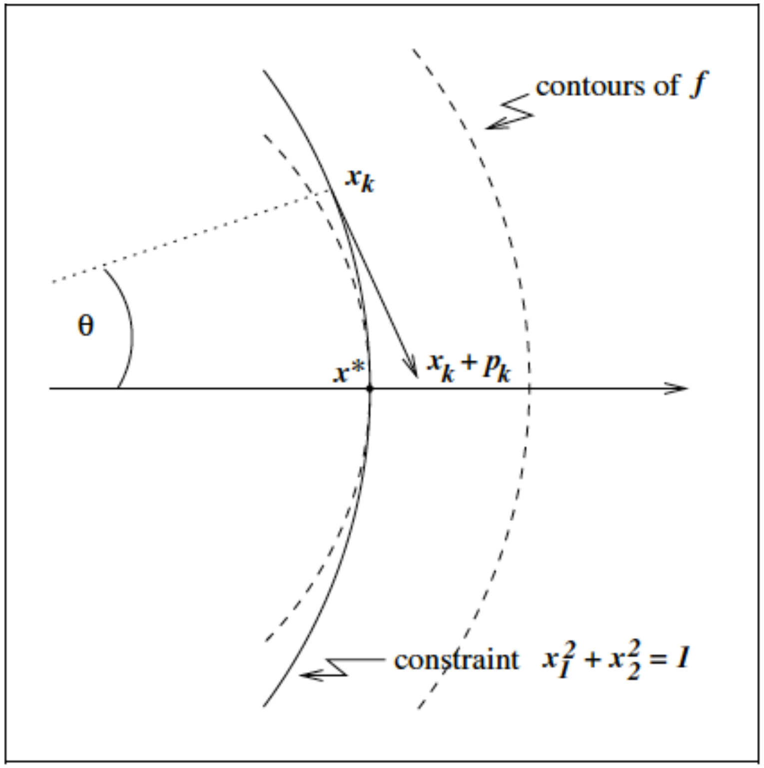
$$f(x^+) \le f_j - \beta h_j \qquad \text{or} \qquad h(x^+) \le h_j - \beta h_j, \qquad \beta \sim 10^{-5} \quad (15.33)$$

# 9.3 MARATOS EFFECT AND CURVILINEAR SEARCH

# Unfortunately, the Newton step may not be compatible with penalty

- This is called the Maratos effect.
- Problem:

$$\min \; f(x_1, x_2) = 2(x_1^2 + x_2^2 - 1) - x_1,$$
$$x_1^2 + x_2^2 - 1 = 0.$$




- Note: the closest point on search direction (Newton) will be rejected !
- So fast convergence does not occur

# Solutions?

- Use Fletcher's function that does not suffer from this problem.
- Following a step: $A_k p_k + c(x_k) = 0.$
- Use a correction that satisfies $A_k \hat{p}_k + c(x_k + p_k) = 0.$

$$\hat{p}_k = -A_k^T (A_k A_k^T)^{-1} c(x_k + p_k),$$

- Followed by the update or line search:

$$x_k + p_k + \hat{p}_k \qquad\qquad x_k + \tau p_k + \tau^2 \hat{p}_k$$

- Since $c(x_k + p_k + \hat{p}_k) = O\left(\|x_k - x^*\|^3\right)$ compared to $c(x_k + p_k) = O\left(\|x_k - x^*\|^2\right)$ corrected Newton step is likelier to be accepted.

# Section 10: Quadratic Programming

Reference: Chapter 16, Nocedal and Wright.

# 10.1 GRADIENT PROJECTIONS FOR QPS WITH BOUND CONSTRAINTS

# Projection

- The problem: $$\min_x \quad q(x) = \tfrac{1}{2}x^T G x + x^T c \quad \text{subject to} \quad l \le x \le u,$$
- Like in the trust-region case, we look for a Cauchy point, based on a projection on the feasible set.
- G does not have to be psd (essential for AugLag)
- The projection operator:

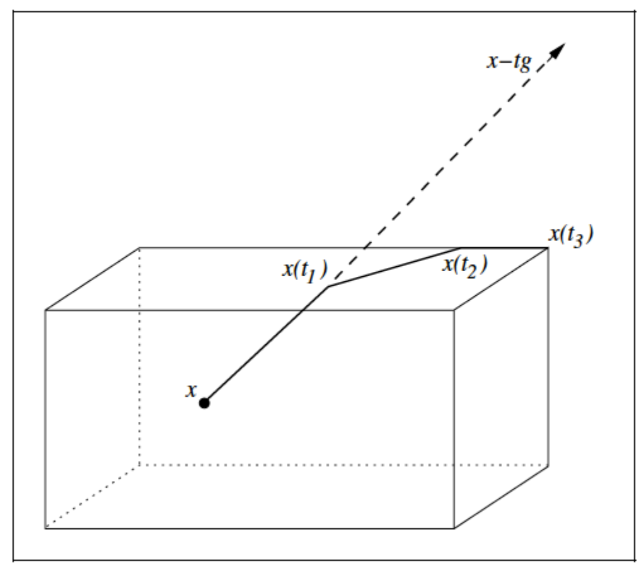
$$P(x, l, u)_i = \begin{cases} l_i & \text{if} \quad x_i < l_i, \\ x_i & \text{if} \quad x_i \in [l_i, u_i], \\ u_i & \text{if} \quad x_i > u_i. \end{cases}$$

# The search path

- Create a piecewise linear path which is feasible (as opposed to the linear one in the unconstrained case) by projection of gradient.

$$x(t) = P(x - tg, l, u),$$

$$g = Gx + c;$$



# Computation of breakpoints

- Can be done on each component individually

$$\bar{t}_i = \begin{cases} (x_i - u_i)/g_i & \text{if } g_i < 0 \text{ and } u_i < +\infty, \\ (x_i - l_i)/g_i & \text{if } g_i > 0 \text{ and } l_i > -\infty, \\ \infty & \text{otherwise.} \end{cases}$$

- Then the search path becomes on each component:

$$x_i(t) = \begin{cases} x_i - t g_i & \text{if } t \le \bar{t}_i, \\ x_i - \bar{t}_i g_i & \text{otherwise.} \end{cases}$$

# Line Search along piecewise linear path

- Reorder the breakpoints eliminating duplicates and zero values to get

$$0 < t_1 < t_2 < \ldots$$

- The path:

$$x(t) = x(t_{j-1}) + (\Delta t)p^{j-1}, \qquad \Delta t = t - t_{j-1} \in [0, t_j - t_{j-1}],$$

- Whose direction is:

$$p_i^{j-1} = \begin{cases} -g_i & \text{if } t_{j-1} < \bar{t}_i, \\ 0 & \text{otherwise.} \end{cases}$$

# Line Search (2)

- Along each piece, $\left[t_{j-1}, t_j\right]$ find the minimum of the quadratic

$$\frac{1}{2}x^T G x + c^T x$$

- This reduces to analyzing a one dimensional quadratic form of t on an interval.
- If the minimum is on the right end of interval, we continue.
- If not, we found the local minimum and the Cauchy point.

# Subspace Minimization

- Active set of Cauchy Point

$$\mathcal{A}(x^c) = \{i \mid x_i^c = l_i \text{ or } x_i^c = u_i\}.$$

- Solve subspace minimization problem

$$\begin{aligned} \min_x \; & q(x) = \tfrac{1}{2} x^T G x + x^T c \\ \text{subject to} \quad & x_i = x_i^c, \; i \in \mathcal{A}(x^c), \\ & l_i \le x_i \le u_i, \; i \notin \mathcal{A}(x^c). \end{aligned}$$

- No need to solve exactly. For example truncated CG with termination if one inactive variable reaches bound.

# Gradient Projection for QP

**Algorithm 16.5** (Gradient Projection Method for QP).

Compute a feasible starting point $x^0$;

**for** $k = 0, 1, 2, \ldots$

  **if** $x^k$ satisfies the KKT conditions for (16.68)

    **stop** with solution $x^* = x^k$;

  Set $x = x^k$ and find the Cauchy point $x^c$;

  Find an approximate solution $x^+$ of (16.74) such that $q(x^+) \leq q(x^c)$

    and $x^+$ is feasible;

  $x^{k+1} \leftarrow x^+$;

**end (for)**

Or, equivalently, if projection does not advance from 0.